自然言語処理 118－1
（1997. 3. 21）

辞書と共起情報を用いた新聞記事からの人名獲得

久光徹* 丹羽芳樹*
日立製作所 基礎研究所
〒350-03 埼玉県比企郡鳩山町赤沼2520
{hisamitu, yniwa}@harl.hitachi.co.jp

要旨

情報検索や情報抽出を目的として新聞記事のように開いたテキストを形態素解析する場合，解析誤りの主因の一つは，辞書に登録されていない，社名，人名，地名等の固有名詞の存在である．本報告では，人名辞書の拡充を目標とし，漢字・片仮名表記の未登録姓名を抽出しつつ，既登録であっても，姓・名に分類されていない固有名詞を，姓・名に分類する方法について述べる．大量の新聞記事からパターンマッチにより抽出した人名候補文字列の集合を長さの順にソートし，既存の辞書と文字列の大域的な出現状況を組み合わせて，姓・名の分割を行う．高い確度で姓または名と推定される文字列が獲得された場合は，その場で辞書に追加することにより，処理の進行とともに獲得精度の向上を図る．新聞記事１年分中，「さん」の前に現れる文字列から抽出した異なり数11,123の文字列を対象とし，約25,000個の人名を含む辞書を用いて行った抽出実験の結果，新たに姓・名893個を約95%の精度で獲得し，既登録の人名3725個の姓・名判別を，約99%の精度で行った．その過程で用いた，人名接辞獲得のための効果的な支援方法についても報告する．

キーワード：コーパス，情報抽出，形態素解析

Acquisition of Person Names from Newspaper Articles by Lexical Knowledge and Co-occurrence Analysis

Toru HISAMITSU and Yoshiki NIWA

Advanced Research Laboratory, Hitachi Ltd.
Hatoyama, Saitama 350-03, Japan
{hisamitu, yniwa}@harl.hitachi.co.jp

Summary

The majority of errors in Japanese morphological analysis is caused by unknown words, most of which consists of proper names such as company names, product names, person names and place names. This paper proposes a method of acquiring unregistered person names from newspaper articles. The method also distinguishes family names from given names. Character strings which are assumed to contain person names are first extracted by pattern matching and sorted in the order of their length. Then each of the strings is divided into a family name and a given name using a lexicon and co-occurrence analysis. A newly found word having enough evidence is immediately added into the dictionary, which increases the accuracy of the following analysis. In an experiment on 11,123 different strings, 893 names were newly acquired with 95.3% accuracy and 3725 registered names were distinguished as family names or given names with 98.5% accuracy. This paper also reports an effective method of acquiring suffixes for person names.

Keywords: corpus-based NLP, information extraction, morphological analysis

1．はじめに

昨今新聞や特許等の大規模な電子化コーパスが容易に入手可能となり，情報検索や情報抽出の対象としてさかんに利用されている．これらの開いたテキストを対象とした形態素解析では，解析誤りの主因の一つは，辞書未登録語，なかでも，人名，地名，社名，製品名等の固有名詞の存在である．

本報告では，人名辞書の拡充のための人名獲得について述べる．第１の課題は，辞書未登録の人名を，姓名の判別をしつつ抽出すること，第２の課題は，姓名の別が記述されていない既登録人名に対して，姓・名の判別を行うことである．

第２の課題は，既存の辞書において，必ずしも固有名詞が細かく分類されていないことが多く，その分類を支援する方法が必要と考えられるため設定した（我々の用いた辞書では，固有名詞の細分類はない）．

今回は第１のステップとして，漢字と片仮名表記された人名を抽出の対象とする（この制限により抽出対象とならない平仮名を含む人名については，その個数の推定，抽出の手法とあわせて，６で述べる）．

未登録語を含んだ形態素解析の研究は，形態素解析研究の初期から現在にかけて，数多くおこなわれており（例えば[1]，[2]），一方で，純粋に統計的な手法で単語を獲得しようとする研究も行われているが[3][4]，これらの研究は，上記の課題に応えるには一般的にすぎる．情報抽出の観点からは，製品名，会社名まで含めた固有名詞同定の枠組みも提唱されているが（例えば[5]），提携情報や製品情報の関連情報として人名を抽出するための枠組みが一般であり，人名に特化した抽出方法を詳細に検討したものではない．従って，本研究は，むしろこれらの研究の要素技術となるものであろう．

本報告では，大量の新聞記事からパターンマッチにより抽出した人名候補文字列の集合から，既存の辞書と，文字列の大域的な出現状況を組み合わせて，姓・名の獲得を行う枠組みについて述べる．高い確度で姓または名と推定される文字列が獲得された場合は，その場で追加辞書に登録し，以降の解析に用いることにより，処理の進行とともに獲得精度の向上を図る．

実験では，新聞記事１年分中，「さん」の前に現れる文字列から抽出した異なり数11,123の文字列を対象とし，約25000個の人名を含む辞書を用いて行った抽出実験の結果，追加辞書に約95%の精度で新たに姓・名を893個を獲得し，辞書登録されていた未分類固有名詞3725個を，約99%の精度で姓・名に分類・識別した．

以下，２では，人名候補文字列獲得パターンの生成について，３では，辞書と組み合わせた人名獲得の手続を述べる．４で実験結果と考察を，５で今後の方針について述べる．

2．人名を含む文字列の獲得

人名を収集するために手がかりとなる単語は，現われ方に関して大きく次の二種類に分けられる：

(I)「さん」，「氏」のように，修飾する人名の直後にあらわれる接辞の集合（S）

(II)「…社長A氏」，「弁護士Bさん」のように，人名の直後のみでなく，同格として直前にあらわれうる自立語の集合(T-I)

(III)「俳優（の）C氏」のように，人名の直前に同格としてあらわれうる自立語の集合（T-II）

以下では，便宜的に，(I)と(II)を拡張人名接辞と呼ぶことにする．

人名の直前・直後における単語の出現パターンは，人名の集合をP，S,P,T-I,T-II以外の単語の集合をCとして，

$c \cdot p \cdot t_1 \cdot c$,

$c \cdot (t_1 \text{ or } t_2) \cdot (の) \cdot p \cdot (s) \cdot c$

($c \in C$，$p \in P$，$t_1 \in$T-I，$t_2 \in$T-II，$s \in S$)

のようになる（括弧内は省略可能）．

Sの要素は，「さん」，「氏」，「ちゃん」など，限られており，新聞記事を対象とした人名抽出のためには，「さん」,「氏」を用いれば実質的に十分である．

漢字，片仮名で書かれた人名の獲得を目標に，「さん」，「氏」の直前から，デリミタと呼ぶ集合（記号，英数字,平仮名等）の要素の手前までの文字列を抽出すると，次のようなものが得られる：

AAA営業本部第１営業部長XXXX
BBB弁護士YYYY
俳優のZZZZ
WWWW
町史編（"町史編さん"の部分文字列）

（AAA，BBB：団体名，XXXX，YYYY，ZZZZ，WWWWは人名）

この集合を以下P_0と呼ぶ．P_0の要素の異

なり数は，46,008であった．

これらから人名部分を抽出するために，我々は拡張人名接辞（網掛け部分）を手がかりとして，上記の文字列から，下線部分を除去することにした．そのためには，これら拡張人名接辞や人名関連自立語を獲得する必要がある．

2.1 パターン記録のためのデータ構造

特定の文字列の前後の文字の出現状況を容易に調べられるように，我々は正・逆方向の頻度付きTRIEを用いた（TRIEについては[6]を参照）．これは，TRIEに複数の文字列を記録するとき，ノード間遷移の回数を文字ごとに計数し，ノード情報に付加したものである．図１に，５つの文字列

{"ABBBB", "ABBBBC", "ABCDE", "ABCDD", "ABC"}

を，正方向の頻度付きTRIEに記録した例を示す．

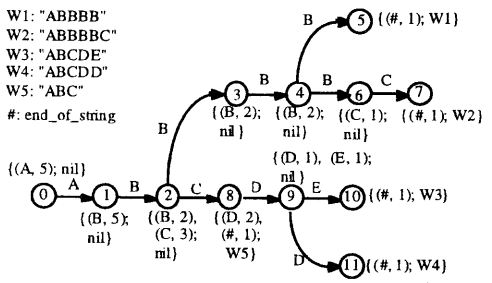

図１頻度付きTRIEの例

2.2 拡張人名接辞等の除去

T-I, T-IIの要素を，内観や目視によって獲得することには限界があるため，これらの獲得を支援する手法を開発した．ここでは，その手法について簡単に述べる．

2.2.1「長」で終わる拡張辞群の獲得

コーパスから，「長」で終わる文字列を，「長」から前方に向かってデリミタの手前まで切りだし，これらを収集したものを集合T_0とする．T_0には，｛社長，営業部長，低成長｝等の文字列がふくまれる．T_0の要素を，逆方向頻度付きTRIEに縮約して記録し，T_0の各要素の，後方から2文字及び3文字を切りだし，これらのT_0中で前方エントロピー[4]を計算し（後述），その値に関してソートする．他の二種類の制約と併用して（詳細は省略）ノイズを除去する．この結果，「低成長」等は除去されるか，低順位を付与され，求めたい文字列の集合が上位に浮上する．このリストから目視により拡張接辞を獲得した．

表１に，目視対象としたリストの先頭20位を，前方エントロピーとともに示す：

1: 社長 10804.03	11: 本部長 3000.74
2: 会長 7120.62	12: 工場長 2804.50
3: 部長 6927.42	13: 所長 2709.88
4: 副社長 4010.68	14: 室長 2544.96
5: 理事長 3585.09	15: 店長 2061.57
6: 委員長 3504.11	16: 副会長 1642.39
7: 課長 3491.54	17: 町長 1483.37
8: 市長 3491.20	18: 裁判長 1455.20
9: 支店長 3195.57	19: 議長 1432.60
10: 局長 3002.40	20: 次長 1073.92

表１

注：文字列sの前に，文字$c_1,\ldots,c_m$が，それぞれ$f_1,\ldots,f_m$回出現したとき，sの前方エントロピーは，

$$\sum_{i=1}^{m} \frac{f_i}{F} \log \frac{f_i}{F}, \quad F = \sum_{i=1}^{m} f_i$$

で定義される．後方エントロピーについても，同様に定義される．

2.2.2 その他の拡張接辞の獲得

「専務」，「弁護士」等の拡張接辞は，特定の文字を手がかりとして収集することが困難なため，頻出人名を媒介に用いる．

step1

P_0の各文字列の先頭から2及び3文字を切りだして，既登録の人名か否かを調べ，人名ならばこれを記録し，P_0中での頻度順にソートして，上位15位までを選ぶ．（この段階で，「鈴木」，「佐藤」などが獲得される）

step2

選ばれた姓で始まる文字列を，デリミタの直前までコーパスから切り出し，これらの集合をT_1とし，逆方向頻度付きTRIEに格納する．

step3

T_1中の文字列の後方から2文字及び3文字を切り出し，P_0中でそれらの前方エントロピーを計算し，その値に関してソートする．

このリストの目視により，「長」で終わらない拡張接辞を獲得した．以下に，目視対

象としたリストの先頭20位を，エントロピーとともに示す：

1: 容疑者 654.53	11: 会頭 102.96
2: 専務 625.51	12: 弁護士 101.56
3: 常務 606.54	13: 選手 90.26
4: 被告 449.92	14: 副知事 78.35
5: 記者 417.47	15: 顧問 74.92
6: 取締役 338.80	16: 製作所 68.48
7: 頭取 163.37	17: 理事 67.09
8: 議員 118.90	18: 長官 63.67
9: 代議士 118.29	19: 事務所 58.20
10: 工業 115.53	20: 医師 58.13

表2

2.2.3 人名関連自立語の獲得

「俳優」，「作家」等は，拡張接辞のような性質を持たないが，人名の直前に現われるため，P_0に混入する．これらを除去するため，これらの単語を次のようにして獲得した：

step1

2.2.1で獲得した拡張人名接辞の集合をT_2とすると，P_0の各要素の，先頭からT_2の要素までを除去する（T_2の要素が複数あるときは，最後尾のものまで．なければなにもしない）．こうして得た文字列の集合をP_1とする）

step2

P_1の要素の先頭側2文字または3文字が既登録の人名でないとき，これらを切りだし，頻度順にソート，目視する．

これにより，｛作家，評論家，無職，高校生...｝等が獲得される．

2.3 先頭側拡張接辞・同格部の除去

2.2に述べた方法で，約300個の拡張接辞および人名関連の自立語を獲得した．これら全体をT_3とすると，T_3を用いて，再びP_1の各要素の先頭からT_3の要素まで（T_2の要素が複数あるときは最後尾のものまで．なければなにもしない）を除去することにより，前方側の接辞等を除去をした文字列を得る．以下，この集合をP_2とする．

P_2の要素の異なり数は，34,145であった．

2.4 参照用データの獲得

後に参照用に用いるため，2.2.2で獲得した拡張接辞のうち30個を選び，これらの直前から，前方のデリミタ手前までの文字列を抽出し，これらをRとする．Rには，「第一営業」のような人名を含まない文字列と，「鈴木」のような人名が混在するが，P_2の要素であって，Rにも含まれるものは，人名である確度が高いため，人名抽出の根拠として用いる．

3．人名の切り出し

2.3までで，「さん」，「氏」の直前から前方のデリミタ手前までの文字列の先頭側から，拡張人名接辞，人名関連自立語を除去した文字列の集合P_2が得られる．P_2の要素一部を示すと次のようである：

（長さ12以上の例は省略）
長さ11:
"江沢民総書記主催歓迎晩",...
長さ10:
"創立七十周年記念晩",....
長さ9:
"博多人形師田中伸幸",...
長さ8:
"大統領歓迎宮中晩", ...
長さ7:
"下部友一社史編", "不動産業沈一男",...
長さ6:
"両陛下歓迎晩", "本名山田美幸",
"保戸塚美枝子", "片岡仁左衛門", ...
長さ5:
"磯村浩之亮", "脇海道弘一", "和田真由美",
"和田崎晃一", "鈴木保奈美", "父親大次郎", ...
...
長さ4:
"薮野清茂", "鞠谷祐子", "圓城三花","脇田純一",
"和辻哲郎", "翌日小松", 福井謙一",...
長さ3:
"筧栄一, "脇海道", "六角彰", "郵政省", "遊亀子",
"朴文淑, "母淳子",...
長さ2:
"和辻", "溥儀", "巫女", "未来", "本屋", ...
長さ1:
"乾", "李", "編", "円", "晩", ...

上で，網掛けを施したものは，姓名でない文字列，もしくは，姓名の前に除去しきれなかった非姓名文字列が結合したものである．

これらは処理精度の低下を招くが，長さ7以上の文字列（全体の1.3%）のほとんどは，除去しきれなかった接辞を含むものであるため，これを対象から除外することに

する．また，長さ1の文字列は，それより長い文字列の解析において，姓・名として分離されない限り，他の姓・名の一部か，独立した姓・名か，ノイズかを特定することが困難であるため，やはり除外する．従って，処理の対象は，P_2の要素のうち，長さ2から長さ6までの文字列である．これを，P_3とする（P_3はP_2の約95%であった）．

3.1 基本的な考え方

3.1.1 姓・名の分割，姓・名判別の根拠

「姓名」は，原理的に人間が恣意的に生成できるものであるため，ある文字列が姓か，名か，混入したノイズかは，実際の用例から判断する他ない．辞書に登録された姓・名は，用例の集積であるが，我々はこれに加えて，人名が含まれる可能性が高い文字列の集合P_2と参照用データRを利用することができるので，これらに支持されるものを，姓・名として推定することにする．紙面の都合上詳細を省くが，基本的な考え方を，例を用いて説明する．

(1) "山田太郎"自体が，既に登録されている（実際には，このような姓または名は無いが）ならば，これは分割しない．

(2) "山田太郎"という文字列が$P_2 \cup R$に少なくとも1回現れ，"山田"と"太郎"が姓名の区別は無くても，人名として登録されているとき，"山田"を姓，"太郎"を名と推定する．

(3) "山田太郎"という文字列が$P_2 \cup R$中に少なくとも1回現れ，"山田"が辞書に人名として登録されており，"太郎"は，辞書には登録されていないが，"太郎さん"のように単独で最低1回出現するか，"鈴木太郎"のように，$P_2 \cup R$中の"山田"以外の要素の直後に最低1回出現するとき，"山田"を姓，"太郎"を名と推定する．

(4) "山田太郎"という文字列が$P_2 \cup R$中に少なくとも1回現れるが，"山田"も"太郎"も辞書に人名として載っていない場合，"山田一郎"のように，"山田"が$P_2 \cup R$中で"太郎"以外の要素の直前に，最低1回出現し，かつ，"鈴木太郎"のように，"太郎"が$P_2 \cup R$中で，"山田"以外の要素の，直後に最低1回出現するとき，"山田"を姓，"太郎"を名と推定する．

つまり，辞書を含む複数の用例に支持されていることを，姓・名切り出しの根拠とする．異なる分割がそれぞれ別の例によって支持されるとき，支持する用例数の多いものを採用する．

これらの条件は，比較的厳しいが，これをクリアしたものを，確度の高い姓・名候補として，追加辞書に登録し，それに続く解析の根拠として用いることにより，処理の効率と精度を向上させる．

3.1.2 2文字の文字列

2文字の文字列は，原則として分割はせず，辞書にも，追加辞書にも登録されていないものについて，姓・名の判断のみを行う．そのためには，P_2中でエントロピー基準を用いる．2文字以上の文字列が実際に姓または名である場合，95%以上の精度で姓，名が判別できる．すなわち，前方エントロピーが大きなものを名，逆のものを姓とするのである．以下は，その例である（数字は，前後エントロピーのうち大きい方）：

竜神 sei 4.75
隆夫 mei 127.56
隆久 mei 8.50
立野 sei 2.42

3.3 手順

以上を，手順としてまとめると，次のようになる．

[I] 長さ順のソート

P_3の要素を，長い順にソートする．これにより，例えば"山田太郎"は必ず"山田一"や，"太郎"より先に処理されるため，"山田"が先に追加辞書に登録されれば，後の解析を効率化できる．

[II] 辞書と文字列の出現状況を併用した姓名分離

長さ6から3までの文字列を対象として，既登録の人名の姓・名判別，未登録人名の確定を行い，充分な支持例があるものを追加辞書に登録しつつ，解析を続行する．

辞書登録する根拠が不足するものでも，姓・名の片方に支持例があれば分割を行い，ファイルに記録する．そうでないものは，リジェクトの印を付ける．

[III] 長さ2の文字列

既登録の姓名と一致するものは分割せず除外する．既登録の姓名で2分割できるもの（"森毅"のような例）も，除外する．残

りをエントロピー基準で姓・名の判定をし，その結果をファイルに記録する．

補助辞書の内容と，[II]，[III]でファイルに記録された姓・名候補は，最終的にはKWIC等を利用して，目視により確定する．

4．実験結果

4.1 対象

我々は，手始めに，P_3の要素のうち,「さん」の前方にあらわれる文字列の集合P_4を用いて実験を行った（これはサイズが手頃であった他には特別な意味はない）．日経新聞の1992年1年分から抽出し，接辞除去処理を加えた結果，P_4の要素の異なり数は11,123個であった．用いた辞書は，エントリー数が約120,000万，固有名は分類されておらず，そのうち人名の推定個数は25,000である．

4.2 追加辞書の精度

処理終了時に，追加辞書には計488個の姓と，405個の未登録人名が獲得され，そのうち正しくないものは42個，精度は95.3%であった．既登録語の中では，3725個の固有名詞を姓・名（名は1612個）と新たに特定し，その精度は98.5%であった．

4.3 考察

4.3.1 エラーの原因

間違いの主因は，"母恵美子さん"の"母"のように，除去されていない非人名部分で，姓のように現れるのもの，"鈴木誠一郎"のような長い名前で，"鈴木誠"と，"一郎"が，それぞれ単独で出現するために，"鈴木誠"と"一郎"に分離されるという現象の二つであった．既登録語の姓・名判別の間違いは，"久保明"→"久"＋"保明"のような，分割誤りであった．

4.3.2 Recallについて

現時点では，精度に関しての詳細な評価は，獲得された辞書のprecisionについてのみしか行っていない．概算ではrecallは50%を越えていると思われるが，辞書登録条件の一部を緩和することにより，precisionをあまり落とさず，recallを向上できると考えている．

5．今後の課題

5.1 平仮名の人名

平仮名を含む人名は，現時点では獲得できないが，数的には，「さん」の直前に300個程度出現していると推定される：これらの大部分は，対となっている漢字の姓を，別の姓名から獲得できると推定されるため，獲得された漢字姓と，「さん」,「氏」にはさまれた部分を抽出することにより，獲得は可能であると考える．

5.2 確率モデルへの展開

姓・名に分離された大量の人名が手に入ると，姓名の分割のための確率モデルを構成することができる．姓・名の区別がわからない場合に比べて，高精度のモデルを構築できると期待される．

6．おわりに

本報告では，姓・名等の細分類がされていない既存辞書と，新聞記事からパターンマッチで抽出した文字列の集合を用いて，未登録の人名を抽出しつつ，既存辞書中の姓・名を特定する方法を述べた．漢字と片仮名表記された人名を抽出の対象とした場合，新聞記事１年分中，「さん」の前に現れる文字列から抽出した異なり数11,123の文字列から，新たに姓・名893個を約95%の精度で獲得し，既登録の人名3725個の姓・名判別を，約99%の精度で行った．

参考文献

[1] 吉村賢治，武内美津乃，津田健蔵，首藤公昭: 未登録語を含む日本語文の形態素解析，情処論文誌, Vol. 30, No. 3, pp.294-300(1989)

[2] 朴哲済，筧捷彦:語の連接関係を利用した未知語の形態素辞書情報の獲得手法，自然言語処理,Vol.4, No.1, pp.71-86(1997)

[3] Nagao,M. and Mori,Y., A New Method of N-gram Statistics for Large Number of n and Automatic Extraction of Words and Phrases from Large Text Data of Japanese, Proc. of COLING'94, pp.611-615(1994)

[4] 下畑さより，杉尾俊之，永田淳次:隣接文字の分散値を用いた定型表現の自動抽出，NL研資料, NL110-11, pp.71-78 (1995)

[5] Kitani, T. and Mitamura, T.: An Accurate Morphological Analysis and Proper Name Identification for Japanese Text Processing, Trans. of IPSJ, Vol.35,No.3, pp.404-413(1994)

[6] Sedgewick, R.: Algorithms, Addison Wesley(1988) pp.248